Genomics 94 (2009) 284–286

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Short Communication

PGA4genomics for comparative genome assembly based on genetic algorithm optimization

Fangqing Zhao [a,b,*], Huabin Hou [a], Qiyu Bao [a], Jinyu Wu [a,*]

[a] Institute of Biomedical Informatics/Zhejiang Provincial Key Laboratory of Medical Genetics, Wenzhou Medical College, Wenzhou 325035, China
[b] Center for Comparative Genomics and Bioinformatics, Department of Biochemistry and Molecular Biology, Pennsylvania State University, Pennsylvania 16802, USA

ARTICLE INFO

Article history:
Received 5 May 2009
Accepted 19 June 2009
Available online 30 June 2009

Keywords:
Bioinformatics
Genetic algorithm
Genome assembly

ABSTRACT

New sequencing technologies greatly facilitate the large-scale bacterial genome sequencing by reducing cost. However, a considerable bottleneck is in the finishing phase, where dozens to hundreds of gaps need to be closed. In this study, we constructed a web server (PGA4genomics) to help users automate gap closing based on comparative genomic syntenies. Extensive evaluations showed that it significantly outperforms previous methods and can produce highly accurate layout result, especially when assembling genomes that are only moderately related. The availability of such a platform would greatly benefit the research community working on bacterial genomics. PGA4genomics can be accessed at two mirror sites http://centre.bioinformatics.zj.cn:8080/pga or http://59.79.168.90:8080/pga.

Introduction

Undoubtedly, the availability of a complete bacterial genome sequence would provide invaluable information for prokaryotic genetics, evolutionary mechanism, structure and function of gene families. It is indicated that, at a minimum, at least one representative for each economically and medically important microbe should have been completely sequenced. Next generation massively parallel sequencing technologies, including Roche 454, Illumina Solexa GA and ABI SOLiD, are able to generate millions of reads in a single run. For example, Illumina Solexa GAII can now offer 75 bp reads and up to 200 million reads per paired-end run. The emergence of these high-throughput technologies has dramatically sped up whole-genome de novo sequencing in a rapid and cost-effective fashion. However, besides the sequencing error, there are two main drawbacks for the de novo genome sequencing: de novo assembly of short reads into contigs and ordering the contigs to generate a complete genome [1]. Recently, several developed tools can make de novo assembly from short reads, and produce dozens to thousands of contigs, including SSAKE [2], Velvet [3], ALLPATHS [4] and EULER-USR [5]. However, the bacterial genome project is still far from finished, because extensive work should be turned to order contigs and to close gaps, which are considered rate-limiting for genome projects.

In our previous study, we proposed a new pheromone trail based genetic algorithm (PGA) to layout contigs [6]. PGA method was demonstrated significantly better than other tools in the following aspects: first, the novel scoring system of evaluating the distance between two contigs can maximize the useful synteny information between the target and reference genomes; second, PGA can distinguish the optimal connection for each contig from misconnections with global search heuristics; third, PGA can get balanced syntenic information from multiple reference genomes. We have successfully applied this algorithm to real incomplete genome data sets produced by Sanger DNA sequencing and pyrosequencing.

However, the usage of PGA is not friendly enough for the users who are not familiar with command-line interface. Non-expert users are then often overwhelmed by the input formats and options on different servers, as well as the difficulty to visualize the layout results. In addition, the users are required to install a number of required bioinformatics packages on the local computer, some of which can be challenging to set up and run, in order to generate the layout results. As requested, we built an integrated online platform (PGA4genomics) for assembling the contigs with one or multiple reference genomes. Pga4genomics is intended to simplify gap closure in genome finishing, by automating contig ordering, orientation, syntenic visualization, and primer design. The availability of such a platform would greatly benefit the research community working on bacterial genomics.

Results and discussion

Input

PGA4genomics provides a simple and user-friendly interface for researchers to generate the layout results (Fig. 1). Users can select a "single" or "multiple" workflow relying on how many reference genomes will be used in the assembly process. In the "single" option, contigs will be ordered based on a single reference genome. PGA4genomics only accepts FASTA format DNA sequences, which can be uploaded from a

* Corresponding authors. Zhao is to be contacted at 312 Wartik Lab, Penn State University, PA 16802 USA. Wu, Institute of Biomedical Informatics, Wenzhou Medical College, 325035, China.
E-mail addresses: fuz3@psu.edu (F. Zhao), iamwujy@yahoo.com.cn (J. Wu).

0888-7543/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2009.06.006

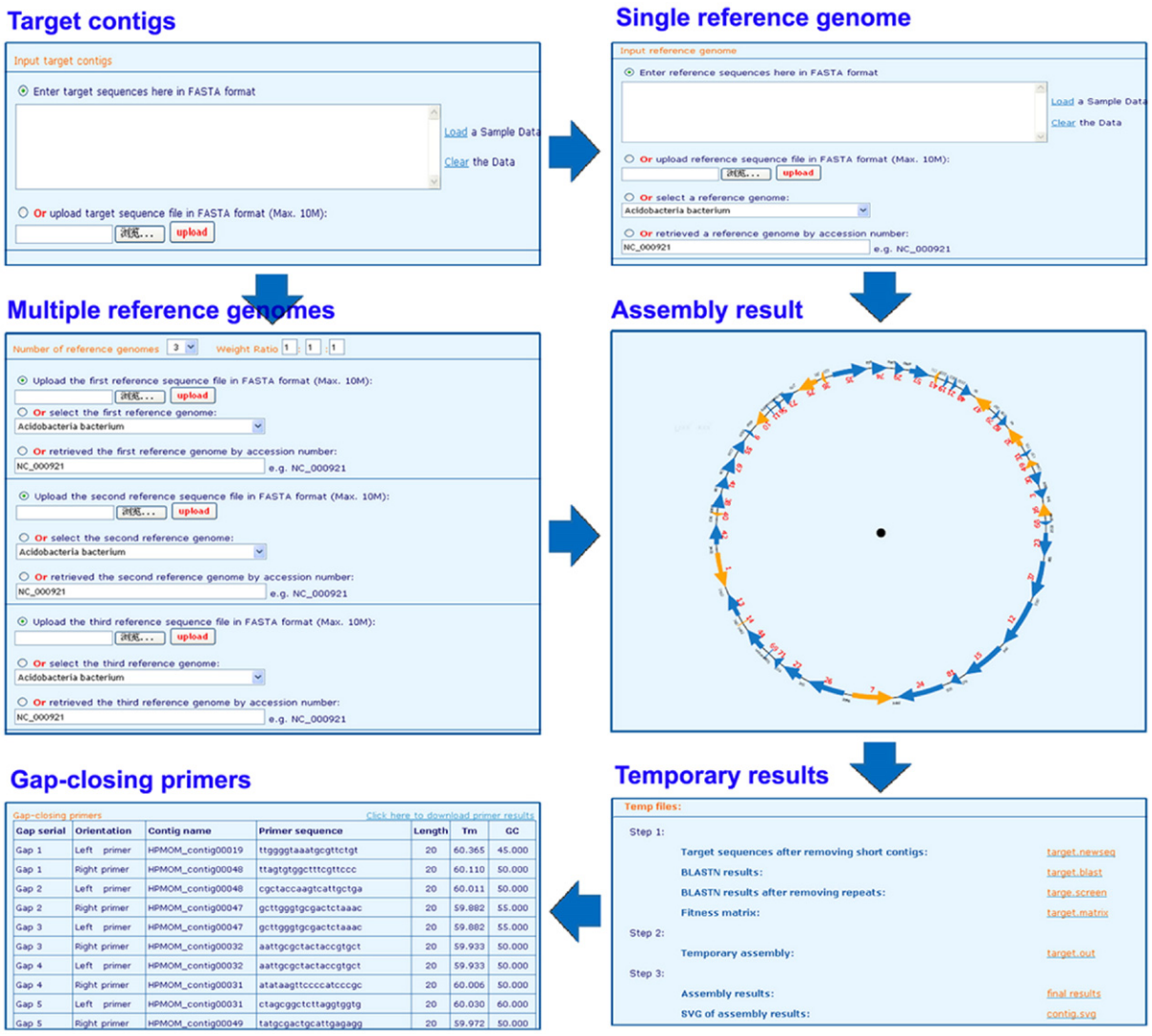

Fig. 1. The user interface and workflow of PGA4genomics.

local disk up to 10 Mb or paste the sequences directly into the input form up to 1 Mb in length. The server will check the validity of the DNA sequence and corresponding size prior to performing any further analysis. Meanwhile, PGA4genomics also provides the sequences for all currently available bacterial genomes. For larger genomes, it would be better to use the Email notification system. Once the job gets done, users will be informed by an Email, which includes a hyperlink to access the results, and the results will be stored on the server for one month.

In the "multiple" mode, users can layout target contigs based on the syntenies identified from multiple reference genomes (Fig. 1). Currently, a maximum of three reference genomes is allowed. Nevertheless, for more reference genomes, users are kindly encouraged to install PGA4genomics in local computers, which is available at http://centre.bioinformatics.zj.cn:8080/pga/download.jsp. It should be noted that the combination of several genomes as the reference may not assure a better result, especially when selecting very closely related and moderately related genomes together, because the latter may obscure connections among contigs. In default, optimal connection for each contig is deduced from the most conserved synteny from specific reference genome. Generally, this default setting may lead to an optimal layout for each contig. Moreover, we also provided a custom weight ratio setting for the input reference genomes, which can tell the algorithm which reference genome is more closely related to the target genome.

A number of important parameters are provided both in the "single" or "multiple" workflow. "Minimal contig size" indicates that only the contigs larger than this threshold are considered; "E-value" is the threshold set for BLASTN; "Match length" shows the minimal length for BLASTN matches; "Iteration number" shows how many iterations PGA will run. Meanwhile, the following three parameters (Beta, Rho, and Q0) have a direct effect on the global search heuristics. The higher Beta-values, the quicker convergence can be found in the early evolutionary stages of the PGA. However, a higher Beta-value usually causes a lower probability of getting the optimal solution. Similarly, the higher Q0 or the lower Rho also results in a quicker convergence at an earlier stage. In default, we set Beta, Rho, and Q0 to 3, 0.8 and 0.8, respectively, because 5000 or 10,000 iterations are sufficient for the evolutionary process to reach convergence, and the final accuracy of the predicted assemblies does not seem to be affected significantly.

Output

When an input form is submitted, a typical run may take several to dozens of minutes to finish, depending on the size of target and reference genomes. The output of PGA4genomics is well organized and consists of three parts: the contig layout map, the temporary files obtained for each step and the gap-closure primers (Fig. 1). The contig layout map is displayed in a SVG format, which is well supported by a

number of commercial and open source softwares, such as CorelDRAW, Illustrilator and ImageMagick. In the map, the arrow indicates the orientation of the contig. The contig that does not have any significant matches in the reference genome is shown in gray color, and is thus randomly located onto the map. The mapping distance between a pair of contigs is labeled on the connected region. If the distance is larger than 10 kb or adjacent to an unmapped contig, it will be indicated with "unknown". In addition, we renamed the contig's name by a sequential number in order to clearly show the layout map. Meanwhile, users can use the zoom function on the middle top of the map when dealing with a large number of target contigs.

In addition to the final layout map, the server also provides several links to the detailed temporary results obtained for each workflow step. The results for the first step include the target sequences after removing short contigs, BLASTN results after removing repeats, and also the fitness matrix. In the second step, five possible assemblies are produced by global search heuristics of the fitness matrix. Because of the nature of genetic algorithm which aims at a near-optimal solution, for each run this algorithm may get a different assembly. To give a better estimation, the default algorithm was set to run five times and thus get five possible assemblies. In the last step, a consensus of above five assemblies is provided, including the optimal contig arrangement, ordering, orientation and distance.

PGA4genomics has also provided a facility to help users design primers for closing gaps. The popular Primer3 program was used to extract optimal primers for each pair of contigs [7], including the following parameters: the fragment size to be scanned by Primer3, spacing from the end of a contig, primer length, primer Tm; product Tm and Primer GC content. Among them, the fragment size to be scanned by Primer3 is set for speeding up the primer design process. The spacing from the end of a contig is the distance between the primer and the terminal of a contig. The primer search results will be displayed in a table format as shown in Fig. 1.

Availability and prospects

In this study, we have developed a web tool PGA4genomics to allow biologists to simplify gap closure in genome finishing by automating contig ordering, orientation, synteny visualization, and primer design. Depending on the number of reference genomes, users can choose the "single" or "multiple" workflow to generate the orientation of target contigs. According to our previous evaluations [6], it significantly outperforms other services, such as Projector2 [8] and OSLay [9], and can produce a highly accurate layout, especially when assembling genomes that are only moderately related. In addition, the server facilitates bacterial genome finishing by providing gap-closure primers. We believe that the simplicity, robustness and accessibility of PGA4genomics can serve as a useful tool in the bacterial genome project to produce high-quality contig layouts. PGA4genomics is free available to all users regardless of academic or non-profit status. There are two available mirror sites http://centre.bioinformatics.zj.cn:8080/pga or http://59.79.168.90:8080/pga.

Materials and methods

Platform algorithm workflow

Three steps are needed to produce the layout results: 1) calculate the fitness matrix; 2) layout contigs based on reference genome; and 3) generate the order and orientation of contigs [6]. In the first step, PGA4genomics workflow begins with retrieving long sequences from FASTA-formatted contigs (target genome), and formatting the user-provided reference genome database. The workflow utilizes BLASTN to detect all possible matches between the target and reference genomes, and subsequently filters repetitive matches. Then the workflow calculates the pairwise distance for each pair of contigs based on above filtered matches. After weighting and rescoring, we finally get a fitness matrix to evaluate possible connections between pairs of contigs. In the second step, the workflow employs global search heuristics to predict the most probable connection for each contig based on the fitness matrix. This global search heuristic ensures that deduced contig pair is the optimal connection for each contig from a collection of possible candidates. In the third step, the contigs can be re-ordered and mapped onto the reference genome, in which the orientation and gap information can be used to design primers and then close gaps.

Platform construction

The associated servlets of the PGA4genomics platform are hosted on an Apache-Tomcat 5.5 server running on Linux operating system. JSP is employed as web engine in web server and Ajax is used to do a form submit without page refreshes. In addition, the layout is displayed in a scalable vector graphics (SVG) format, which requires SVG plug-in to be installed in the client's computers. The web interface of PGA4genomics is implemented in an operating-system independent way and has been tested in Internet Explorer 6.0, Firefox 3.0.1, and Opera 10.00 browsers.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (30800643) and Foundation of Zhejiang Provincial Top Key Discipline of Laboratory Medicine, China.

References

[1] J.M. Aury, C. Cruaud, V. Barbe, O. Rogier, S. Mangenot, G. Samson, J. Poulain, V. Anthouard, C. Scarpelli, F. Artiguenave, P. Wincker, High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies, BMC Genomics 9 (2008) 603.
[2] R.L. Warren, G.G. Sutton, S.J. Jones, R.A. Holt, Assembling millions of short DNA sequences using SSAKE, Bioinformatics 23 (2007) 500–501.
[3] D.R. Zerbino, E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs, Genome Res. 18 (2008) 821–829.
[4] J. Butler, I. MacCallum, M. Kleber, I.A. Shlyakhter, M.K. Belmonte, E.S. Lander, C. Nusbaum, D.B. Jaffe, ALLPATHS: de novo assembly of whole-genome shotgun microreads, Genome Res. 18 (2008) 810–820.
[5] M.J. Chaisson, D. Brinza, P.A. Pevzner, De novo fragment assembly with short mate-paired reads: does the read length matter? Genome Res. 19 (2009) 336–346.
[6] F. Zhao, F. Zhao, T. Li, D.A. Bryant, A new pheromone trail-based genetic algorithm for comparative genome assembly, Nucleic Acids Res. 36 (2008) 3455–3462.
[7] S. Rozen, H. Skaletsky, Primer3 on the WWW for general users and for biologist programmers, Methods Mol. Biol. 132 (2000) 365–386.
[8] S.A. van Hijum, A.L. Zomer, O.P. Kuipers, J. Kok, Projector 2: contig mapping for efficient gap-closure of prokaryotic genome sequence assemblies, Nucleic Acids Res. 33 (2005) W560–W566.
[9] D.C. Richter, S.C. Schuster, D.H. Huson, OSLay: optimal syntenic layout of unfinished assemblies, Bioinformatics 23 (2007) 1573–1579.